It is ordered that the Jury Composition Rule is hereby adopted by this Court, effective **July 1, 2012**, to read as follows:

JURY COMPOSITION RULE

1. Purpose. The purpose of the rule is to set reasonable standards for the preparation, dissemination and improvement of inclusive statewide and county master jury lists.

2. Business Rules. The statewide and county master jury lists shall be compiled substantially in accordance with the business rules set forth in Appendix A.

3. Inclusiveness. Each county master jury list should be no less than 85% inclusive of the number of persons in the county population age 18 years or older as derived from the most recent decennial census or county population estimate (Table B01001 as of the date of this rule) from United States Census Bureau for the calendar year when the list is generated. The calculation shall be made by dividing the number of persons in such master list by the county population age 18 years or older according to the applicable census data. In the event that such percentage is less than 85%, the Council of Superior Court Clerks will provide the county data collected pursuant to OCGA § 15-12-40.1 and applicable census data so that the chief judge may make a prima facie determination whether the list is fairly representative based upon:[1]

a. The findings of the Georgia Supreme Court in representativeness challenges;
b. The level of representativeness; and
c. The alternatives available to increase the inclusiveness of the list.

4. Certification.

a. Upon completion of the statewide and county master jury lists, the Council of Superior Court Clerks or its list vendor shall certify to the Supreme Court that it has complied with the business rules for preparation of the master jury list and that the county master jury lists do or do not meet the inclusiveness threshold.

b. The Council of Superior Court Clerks or its list vendor shall provide written certification of the county master jury list to each county after payment of the subscription invoice presented to the county in conjunction with the delivery of the county master jury list as provided by OCGA § 15-12-40.1. This certification shall include:

i. The year the list was created;
ii. The name of the county;
iii. Certification that the business rules established by this court rule have been followed; and
iv. The percentage inclusiveness of the county master jury list as certified to the Supreme Court.

5. The written certificate shall be provided to the trial court and shall be included in the trial judge's report as required by OCGA § 17-10-35 (a).

---

[1] See National Center for State Courts, Trial Court Performance Standards & Measurement System, Standard 3.2.3: Representativeness of Final Juror Pool (last modified January 2005).

6. Local clerks and jury commissioners shall not add or delete names from the county master jury list, but may excuse, defer, or inactivate names of jurors known to be ineligible or incompetent to serve pursuant to OCGA § 15-12-1.1. The clerk of the board of jury commissioners shall maintain a list of jurors excused, deferred or inactive who are not part of the eligible juror array derived from the county master jury list.

7. All other issues of local jury management shall be as authorized by law or by local court order.

8. In the promulgation of this rule, the Court does not express any advisory opinion on the legal sufficiency of compliance.

APPENDIX A: INCLUSIVE SOURCE LIST: PROCESS AND BUSINESS RULES

PRIMARY RECORDS SOURCES

The following shall be used as the two sources of data for the creation of the statewide and county master jury lists. Such sources are hereafter referred to as "Primary Records Sources."

Department of Driver Services (DDS)

Records shall be secured from the Georgia Department of Driver Services (DDS). Such records shall include data relating to all persons 18 years of age and older with any of the following:

(a) valid and expired driver's licenses,
(b) state issued personal identification documents, or
(c) records of in-state and out-of-state convictions for driving without a license, revocations, and suspensions.

Secretary of State Voter Registration Records

Voter registration records shall be secured from the Georgia Secretary of State. Such records shall include data relating to all persons registered to vote within the state, including persons identified by the Secretary of State as "active" and "inactive."

LIMITING RECORDS SOURCES

The following record sources shall be used as sources of data to be applied to the Primary Records Sources to purge persons from the Primary Records Sources as indicated:

Department of Public Health Death Certificates

Death certification data shall be obtained from the Department of Public Health including data relating to all current and past (15 years) Georgia death certificates. The certificates include first name, last name, middle name, gender, date of birth, address/county of death, and county/address of residence.

Records shall be purged from the Primary Records Sources relating to all persons found in the death certificate file when such records match on each of the six fields stated below. Matching shall be made using deterministic matching methods and the following fields:

1. County of Residence
2. Last Name
3. First Name (or use first four characters of name)
4. Middle Initial
5. Sex
6. Date of Birth

<u>Secretary of State: List of Convicted Felons</u>

A list of persons shall be obtained from the Secretary of State for all persons who have been convicted of felonies in state or federal courts and who have not had their civil rights restored. Felons shall be purged from the merged source file.

<u>County Permanent Excusals</u>

A request shall be made of each Superior Court Clerk or county jury clerk for an electronic listing of all persons within such county who have been permanently excused or inactivated from jury service as follows:

(a) such persons who have been permanently excused or inactivated due to mental and/or physical disability; and

(b) such persons who are 70 years of age or older and who have requested and been granted permanent excusals or inactivation from jury service as the result of their age.

Such listing shall include such data elements as specified by the Council of Superior Court Clerks. Such listings shall be submitted by reasonable deadlines as determined by the Council.

Persons appearing on such lists presented in a timely manner shall be inactivated from the county master jury list prior to delivery to each county; to the extent that local listings are not timely submitted to the Council, the Council shall still provide a county master jury list.

This provision shall not limit the authority of the court to excuse or inactivate such persons locally.

<u>Source List Preparation and Business Rules</u>

Compiling the sources is conducted sequentially after receiving the DDS, voter registration, and death certificates. The following sections provide specific business and process description.

Data Filters

Prior to standardizing and clearing the eligible source list records, the first step is to purge records from the DDS source data. The specific business rules guide pre-merge record purging. Six fields are used to purge ineligible DDS records:

1. License Status
2. Personal ID Flag indicating (1) License or (2) Personal State Issued Identification
3. DDS Driver's License #
4. Address Date
5. Date of Document Expiration (License/ID)
6. DDS extraction date

The DDS data extraction date is not included as a data field but is needed to filter expired licenses.

DDS Source Data Filter Rule #1:

Purge from the DDS data any record where the License Status equals “No License” and the Personal ID/Licenses Flag equals “License.” Do not purge records where the Personal/ID field equals “I”.

DDS Source Data Filter Rule #2:

Purge from the DDS data any record where:

(a) License Status = Expired and days since the expiration date is greater than 730 days, and
(b) Personal ID/License Field = “L” (License)

DDS Source Data Filter Rule #3:

Do not purge from the DDS data any record of a state-issued ID even if it appears expired.

DDS Source Data Filter Rule #4:

Purge from the DDS data duplicate record(s) when two or more records have the same Driver's License ID #. The single record retained shall be the record containing the most recent:

(a) address date, or
(b) expiration date, or
(c) document issue date.

Voter Registration Filter Rule #5:

No filters are applied to the voter registration records (inactive voters remain in the final list).

Address Standardization and Cleaning

Name and address standardization procedure shall be performed prior to submission to the National-Change-of-Address (NCOA) vendor or the vendor can authorize NCOA vendor to perform these data cleaning services.

1. Apply software algorithms to extract, parse, and standardize voter/driver address from text fields to ensure the address is consistent with the national United States Postal Service (USPS) Address Information System (Postal Addressing Standards Publication # 28, April 2010).

2. Standardized addresses are matched to the USPS Address Information System to identify potentially invalid addresses. Invalid addresses shall be identified but shall be retained.

3. If the address is missing a ZIP code or has the wrong county code, the USPS Automated Address System is used to correct address components if possible (5-digit ZIP Code, add 4-digit ZIP Code suffix, correct county code).

4. Although voter/driver records have separate first, last, and middle data elements, standardization algorithms standardize special cases (hyphenation, apostrophes).

5. Ensure Georgia DDS county codes correspond to Georgia voter registration county codes. Assign the Federal Information Processing Standards (FIPS) codes to all records (required in subsequent steps to reconcile NCOA returns to DDS/Voter county codes).

National-Change-of-Address (NCOA) Processing

NCOA Rule #1: NCOA Service

1. The NCOA vendor must use the 48-month USPS NCOA database.

2. The selected NCOA vendor must do all processing in-house and cannot outsource any or part of the DDS or voter file matching to other companies or entities.

NCOA Rule #2: NCOA Service

The NCOA vendor shall report whether a residence move is an out-of-state, intra-county, or inter-county move. All records indicating out-of-state moves shall be purged. All records indicating corrected intra-county moves shall be retained.

NCOA Rule #3: NCOA Service

Keep all records even if the NCOA match to the USPS valid address database flags the record as invalid. This USPS address validity flag will be retained in the master source list for clerks to verify accuracy over time using manual checks or returned jury summons. After one year, the clerks can evaluate the accuracy of the undeliverable flag to determine whether to purge these records prior to compiling the list.

Identifying Duplicate Records

Apply "Probability Linking Methods" as described below.

Unlike the deterministic approach which requires an exact match on some or all fields, Probability Records Linkage (PRL) methods use the statistical properties of a record pair to calculate the probability that the records apply to the same person. Exact matches on all the fields are therefore not required. The PRL method allows for both agreements and disagreements among matching fields between two records. PRL takes into account the probability that the matching field, such as the birth month, agrees by chance alone, even if the record pair is not the same person.

For example, suppose birth month is used as one of the matching fields. What is the chance that any pair of records from the voter and driver's license files will have same birth month, even if the two records are not the same person? For the sake of simplicity, let's say that there is an 8% (1/12) chance of agreement on birth month by chance alone, even if the records belong to different people. The power behind PRL becomes more apparent when using a combination of matching fields, such as the surname. The somewhat unusual name "Wilenski" will carry a much higher matching weight than "Smith," which is a very common name. For both the voter and driver's license databases, the frequencies (probabilities) are computed or each value in each of the matching fields. When all agreements and disagreements among these fields and their corresponding weights are computed for each record pair, it is possible to make statements as to the likelihood that the record pair in fact represents the same person.

Identifying Duplicate Records: Methodology

The PRL methods to be used rely on the Fellegi-Sunter (1969) framework to compute odds ratios (see Section 2.1 in the attached article) and a limited Bayesian Model (see Section 3.1). The matching methodology does not apply the full Bayesian Model as described in Section 4.1.

Although the Fellegi-Sunter framework will provide an odds-ratio, it is very difficult to identify an optimal cutting point in terms of successful and unsuccessful matches without manual review. Additionally, odds-ratios do not translate easily into practical interpretation, making it difficult to describe record matching success.

For this reason, the limited Bayesian Model shall be used to convert likelihood ratios (match weights) by converting these estimates into Bayesian posterior probabilities. The limited Bayesian formulae permit computation of an actual probability stating the likelihood that the record pair is indeed a link.

For a complete description, see the attached article: McGlincy, *A Bayesian Record Linkage Methodology for Multiple Imputation of Missing Links*. The references in this article also provide the citations for the supporting matching research (Fellegi-Sunter, Newcombe, and Winkler).

Identifying Duplicate Records: PRL Model Parameters

Blocking Fields:

1. County
2. Gender
3. Last Name (Soundex)
4. Year of Birth

Matching Fields:

1. Last Name
2. First Name (with one typo permitted)
3. Middle Name (first three characters)
4. Birth Day
5. Birth Month
6. Birth Year

Probability Level: 90% or higher

Identifying Duplicate Records: Selecting the DDS or Voter Registration Records between Two Linked Records

Among record pairs that meet or exceed the 90% probability level, the following business rules are used to select the record (DDS or voter) with the best information. In most cases, the voter registration record will have the most recent and complete data in terms of street address so the voter registration record will be selected as the primary record among duplicates. However, this may not always be the case. If so, the following rules apply.

1. Conduct a field-to-field comparison between the two linked records to identify missing data and inconsistent data, such as different addresses.

2. Use DDS address-change date and/or date of license issue and compare these dates to the address and voter date-of-last-contact date. The source record with the most-current dates will dictate what address is used as the selected address.

The statewide master jury list and the county master jury lists shall contain at least the following fields:

1. Last Name
2. First Name

3. Middle Name
4. Birth Day
5. Birth Month
6. Birth Year
7. Residence Address (including City, ZIP Code and County)
8. Mailing Address (including City, ZIP Code and County) if not the same

A Bayesian Record Linkage Methodology for Multiple Imputation of Missing Links

Michael H. McGlincy

Strategic Matching, Inc.
PO Box 334, Morrisonville, NY 12962
Phone 518 643 8485, mcglincym@strategicmatching.com

ABSTRACT

Probabilistic record linkage can be an effective research technique even if available records lack strong personal identifiers or if identifying fields contain many errors or omissions. Traditional methodologies typically select a single set of linked record pairs for research based on a match weight test statistic and clerical review of marginal pairs. However, missing links (false negatives) can make such datasets unrepresentative of the total population of true linked pairs. The methodology described here addresses this problem. A full Bayesian model is developed for the posterior probability that a record pair is a true match given observed agreements and disagreements of comparison fields. Observed-data posterior distributions for model parameters and true match status are estimated simultaneously through MCMC data augmentation with parallel chains. This process gives multiple complete representative sets of imputed linked record pairs. Population estimates can be obtained from each imputation and consolidated using established techniques. Application of linkage imputation by a consortium of traffic safety researchers is described.

KEY WORDS: Bayesian, Record Linkage, Multiple Imputation

1. BACKGROUND

The National Highway Traffic Safety Association (NHTSA) supports the Crash Outcome Data Evaluation System (CODES) program in order to learn about medical consequences of motor vehicle crashes. CODES grantees in 30 states link police crash reports to medical treatment records for all crashes in those states and all injured vehicle occupants. Unique identifiers for persons and events are not available in most CODES datasets. Consequently, CODES researchers use probabilistic record linkage techniques to create linked datasets for analysis (Jaro, 1995; Runge, 2000).

Police crash reports are linked to ambulance run reports from EMS agencies, emergency department or inpatient treatment records from hospitals, or death records from a state vital statistics office (McGlincy *et al.*, 1994; Vernon *et al.*, 2004). Most crash records do not link to treatment records because most occupants are uninjured. Most treatment records do not pertain to crash victims because there are many other reasons for EMS or hospital treatment. CODES researchers carry out frequency-based linkage procedures using commercial software available for that purpose (Jaro, 1992; McGlincy, 2003). Linking such records can be problematic: Datasets of interest usually do not include strong personal identifiers or identifiers are not made available in order to protect patient confidentiality. Furthermore, data quality may be degraded by high levels of nonresponse or misreporting (Greenberg, 1996). Analysts can never be certain about the true match status of any pair of records.

Fellegi and Sunter (1969) suggest a theoretical framework for record linkage under such conditions of uncertainty. In principle, analysts select a single set of linked record pairs that can be treated as the set of all true matched pairs for all practical purposes. In practice, the true disposition of many record pairs might be apparent only after detailed clerical review of information not captured in a computer file or coded in a record linkage model. When clerical review is not feasible because of the lack of identifying data or other program limitations, resulting linked datasets can be characterized by few false positives but many false negatives, or missing links. Those record pairs which happen to have high likelihood of being true matches may not be representative of the total population of true matches. For example, if crashes in rural areas involving elderly drivers are less common than crashes in urban areas involving young drivers then record pairs for rural elderly would be assigned higher likelihoods in frequency-based linkage models than pairs for urban young. This issue is of particular interest to the consortium of CODES researchers because characteristics of available data vary substantively from member to member. Analysis results based on linked datasets from different members cannot easily be compared or combined in a meta-analysis unless they are all representative samples from underlying populations.

The problem of missing links is similar to the problem of nonresponse in surveys. Cases which happen to have complete data may not be representative of the underlying population. Bayesian multiple imputation is a standard technique for analyzing incomplete data (Little and Rubin, 2002; Schafer, 1997). Multiple imputation is used to correct for missing data in CODES datasets and in other NHTSA programs (Rubin, *et al.*, 1998). Here we treat the true match status of all pairs of records as missing data and use the iterative technique of Markov Chain Monte Carlo (MCMC) data augmentation to draw from a Bayesian posterior distribution for the missing match status. We also draw from posterior distributions for parameters of the linkage model.

In Section 2 we summarize the frequency-based linkage model that serves as the framework for CODES linkage projects. An optimal linkage rule based on likelihood ratios is used to select record pairs for analysis. In Section 3, we describe a feasibility test of Bayesian linkage imputation with CODES datasets. CODES researchers in ten states conducted test linkage projects and obtained results that suggest linkage imputation can correct for missing links. In the test, a limited Bayesian model was used to estimate posterior probabilities for a true match given the likelihood ratios described earlier. In Section 4, we extend the methods described in Sections 2 and 3 while maintaining continuity with prior CODES work by developing full Bayesian models for estimating required population characteristics and the true match status of comparison pairs. In Section 5, we describe areas for future work.

Others researchers have assumed different record linkage models (see, for example, Belin and Rubin, 1995; Fortini *et al.*, 2001, 2002; Larsen, 1999 and 2003; Larsen and Rubin, 2001; Winkler 1988, 1989, 1993, 1994). Most of these other models consider two comparison outcomes: agree or not agree, where the latter outcome includes missing values. The frequency-based model described here considers a broader set of comparison outcomes: agreement on specific values, disagreement, or missing. Furthermore, the model includes misreporting and is easily extended to linking three or more files.

Misreported data can introduce bias by attenuating estimates of associations between explanatory variables and outcome variables (Gustafson, 2000). False positive matches (incorrect links) act like misreported data in this respect. This is of particular concern with linkage imputation because each imputed dataset necessarily includes a higher level of false positives. Methods for correcting this bias have been proposed (Lahiri and Larsen, 2000; Larsen, 2003; Scheuren and Winkler, 1993, 1997). This issue is not considered further here.

2. RECORD LINKAGE WITH AN OPTIMAL LINKAGE RULE

2.1 Record Linkage Framework

Fellegi and Sunter (1969) suggest a general theoretical framework for record linkage. Computer records in two files, L_A and L_B, are generated as samples from two populations, A and B, respectively. The problem is to identify those pairs of records pertaining to the same individual. Such record pairs are called *matched* and all other record pairs are called *unmatched*. Recorded characteristics on pairs of records are compared. Comparison pairs are classified according to a decision rule as a *link* if the records are probably for the same individual, a *non-link* if the records are probably not for the same individual, or a *possible link* if there is not sufficient evidence for a positive classification at specified error levels. An optimal linkage rule minimizes the need for clerical review of possible links. The optimal rule ranks comparison pairs by a test statistic m/u, where m is the probability of observing a given comparison outcome on a matched pair and u is the probability of observing the same outcome on an unmatched pair. $\text{Log}_2(m/u)$ is called a *match weight*. Pairs with weights above a cutoff value are classified as links. Pairs with weights below a second, lower cutoff are non-links. Pairs with weights between the cutoff values are possible links.

2.2 Practical Implementation

Fellegi and Sunter (1969) suggest simplifications for practical implementation of their theory. Comparison fields are conditionally independent on the sets of matched and unmatched pairs. Comparison outcomes are limited to agreements on a specific value, disagreements, or missing values. Prior knowledge is assumed about each comparison field. For concreteness, suppose that one comparison field in files L_A and L_B is a person's age. Let p_A, p_B, and p_{AB} be probabilities of observing specific ages in populations A, B, and A∩B, respectively. Let e_{A0} and e_{B0} be probabilities for missing ages for populations A and B, respectively. Let e_A and e_B be probabilities for misreported ages for populations A and B, respectively. Let e_T be the probability of correct ages being reported differently for population B than for population A. Method I gives the following rules for calculating m and u probabilities, given the model parameters (Fellegi and Sunter, 1969, pp. 1192–1193).

$$m\text{ (age agrees and is the } j^{th} \text{ listed age)} = p_{ABj}\,(1-e_A)(1-e_B)(1-e_T)(1-e_{A0})(1-e_{B0})$$

$$m\text{ (age disagrees)} = [1-(1-e_A)(1-e_B)(1-e_T)](1-e_{A0})(1-e_{B0})$$

$$m\text{ (age missing on either file)} = 1-(1-e_{A0})(1-e_{B0})$$

u (age agrees and is the j^{th} listed age) =
$p_{Aj}\, p_{Bj}\, (1 - e_A)(1 - e_B)(1 - e_T)(1 - e_{A0})(1 - e_{B0})$

u (age disagrees) =
$[1 - (1 - e_A)(1 - e_B)(1 - e_T)\, \Sigma_j\, p_{Aj}\, p_{Bj}](1 - e_{A0})(1 - e_{B0})$

u (age missing on either file) =
$1 - (1 - e_{A0})(1 - e_{B0})$

Fellegi and Sunter note that under appropriate conditions the proportions p_{ABj}, p_{Aj} , and p_{Bj} may be estimated from the data files themselves. CODES linkage projects use maximum likelihood estimates of population proportions.

3. A PROOF OF CONCEPT FOR LINKAGE IMPUTATION

3.1 A Limited Bayesian Model

The current CODES linkage methodology includes a limited Bayesian model so that CODES researchers can compare the quality of their linked record pairs in terms of posterior probabilities rather than likelihood ratios or match weights. The model is specifically designed to provide the same ranking of comparison record pairs using any of these measures.

For a given record pair (a,b) and comparison result vector γ for (a,b), let M be the hypothesis that the pair is matched and U that the pair is unmatched. The set of all observed comparison vectors $\Gamma_{AB} = \{\gamma(a,b),\ (a,b) \text{ in } L_A \times L_B\}$ can be considered as arising from a mixture of comparison vectors from two classes, matched and unmatched pairs (Larsen , 1999), so that

$$P(\gamma(a,b)) = P(\gamma(a,b) \mid M)\, P(M) + P(\gamma(a,b) \mid U)\, P(U).$$

The observed-data likelihood for the set of all observed comparison vectors is

$$P(\Gamma_{AB}) = \textstyle\prod_{(a,b)} P(\gamma(a,b)).$$

We assume the latent class or conditional independence model (Larsen and Rubin, 2001) in which the comparison outcome for any field k on record pair (a,b) is independent of the comparison outcomes for other fields. In this case, the conditional probability of observing comparison vector $\gamma(a,b)$ given class H = M or U is the product of K independent conditional probabilities for the comparison outcomes for each of K comparison fields

$$P(\gamma(a,b) \mid H) = \textstyle\prod_k P(\gamma_k(a,b) \mid H),\ H = M \text{ or } U.$$

For Bayesian inference about true status of pair (a,b), we apply Bayes' rule for odds (Gelman *et al.*, 1995, pp. 7–10): the posterior odds for M given γ are the product of the likelihood ratio and the prior odds.

$$\begin{gathered}\text{Odds}(M \mid \gamma(a,b)) \equiv P(M \mid \gamma(a,b)) / P(U \mid \gamma(a,b)) = \\ (P(\gamma(a,b) \mid M) / P(\gamma(a,b) \mid U))\, (P(M) / P(U)) = \\ (\textstyle\prod_k P(\gamma_k(a,b) \mid M) / \prod_k P(\gamma_k(a,b) \mid U))\, (P(M) / P(U)) = \\ (\textstyle\prod_k P(\gamma_k(a,b) \mid M) / P(\gamma_k(a,b) \mid U))\, (P(M) / P(U)) = \\ (\textstyle\prod_k m_k(a,b) / u_k(a,b))\, (P(M) / P(U)).\end{gathered}$$

Conditional probabilities $m_k(a,b)$ and $u_k(a,b)$ are calculated for each pair (a,b) and each field k = 1, …, K given $\gamma(a,b)$ using the rules in Section 2 just as in a non-Bayesian linkage. Given posterior odds for M, the corresponding posterior probability is

$$P(M \mid \gamma) = \text{Odds}(M \mid \gamma) / (1 + \text{Odds}(M \mid \gamma)).$$

For a non-imputed linkage, all high probability pairs are selected for analysis, say $\text{Odds}(M \mid \gamma(a,b)) > 9$ or $P(M \mid \gamma(a,b)) > 0.9$.

We assume an informative prior distribution for the odds based on substantive data: the number of records in the samples L_A and L_B and the number of records in $L_A \cap L_B$ (i.e., true matched pairs). This is feasible for CODES linkages (and many others) because datasets of interest include information about which records should match. For example, hospital records include codes which indicate whether a patient's injuries resulted from a motor vehicle crash. We assume no uncertainty about the prior odds so that a point estimate suffices:

$$\text{Odds}(M) = N_M / N_U = N_M / (N_A\, N_B - N_M),$$

where N_M is the number of matched pairs, N_U is the number of unmatched pairs, N_A is the number of records in A, and N_B is the number of records in B.

Point estimates are assumed for all components of the model parameter $\theta = (p_A, p_B, p_{AB}, e_{A0}, e_{B0}, e_A, e_B, e_{AB})$. Consequently, imputed linkages do not reflect uncertainty in θ. Population proportions p_A, p_B, e_{A0}, and e_{B0} for each field are set equal to observed sample proportions, the MLE values. p_{AB}, e_A, e_B, e_{AB} for each field are set based on prior knowledge and adjusted after inspection of those record pairs classified as matched in a preliminary linkage.

3.2 A Proof of Concept

In order to test the potential of Bayesian imputation for finding missing links, selected researchers conducted the same linkage project following two different methodologies. First, researchers identified the set of high probability links ($P(M \mid \gamma) > 0.9$) between police crash reports and hospital discharge records as usual. Second, researchers obtained five imputed sets of links between the same two datasets. For each imputed linkage, a uniform random deviate X in (0,1) was generated for each comparison pair and the pair was included the imputation if $P(M \mid \gamma) > X$.

3.3 Linkage Imputation Results

Figure 1 illustrates typical linkage imputation results. This figure shows a histogram of match weights for Imputation 1 for the test linkage projects conducted by one CODES state. Here a linked pair with match weight near 22.4 has a posterior probability near 0.9 and the histogram interval for the mode is 19 to 21. High weight counts are essentially the same from imputation to imputation because most high probability links are drawn in most imputations. Low weight counts show more variation because most low probability links are drawn in at most one imputation.

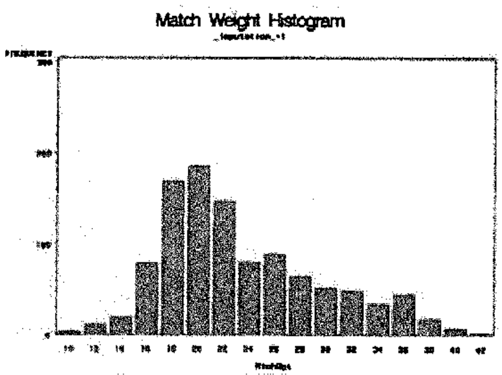

Figure 1. Match Weight Histogram
for Imputation 1

3.4 Comparison of Linkage Results

As shown in Table 1, CODES researchers reported that the percent of linked pairs found that had posterior probabilities greater than 0.9 was consistently less than the estimated total number of pairs based on prior knowledge (41% to 70% of estimate). The number of imputed linked pairs was approximately equal to the estimated number (91% to 111% of estimate), suggesting that Bayesian linkage imputation can account for all missing links.

Table 1
Number of Linked Pairs Found
as a Percent of Prior Estimates

CODES State	*Percent of Prior Estimate*	
	Post. Prob. > 0.9	*Imputation 1*
A	41%	91%
B	56%	111%
C	46%	94%
D	56%	91%
E	55%	105%
F	70%	95%
G	44%	108%
H	69%	97%
I	59%	92%
J	61%	106%

4. A FULL BAYESIAN MODEL FOR RECORD LINKAGE

4.1 The Full Model and Data Augmentation Procedure

This full Bayesian model corrects shortcomings in the limited model used for the proof of concept. Let $\theta = (p_A, p_B, p_{AB}, e_{A0}, e_{B0}, e_A, e_B, e_{AB})$ be a vector parameter consisting of all of the unknown probabilities needed for the model described in Section 2. Let Y_{MAT} be a vector indicating the missing true match status, 1 = matched and 0 = unmatched, for each record pair (a,b) in the Cartesian product $L_A \times L_B$. Let $Y_{OBS} = \{L_A, L_B, \Gamma_{AB}\}$ be the observed data where L_A and L_B are representative samples of records from populations A and B, respectively, and Γ_{AB} is the set of comparison vectors $\gamma(a,b)$ for each record pair (a,b) in $L_A \times L_B$. Denote the posterior distribution for Bayesian record linkage as

$$P(\theta, Y_{MAT} \mid Y_{OBS}) =$$
$$P(p_A, p_B, p_{AB}, e_{A0}, e_{B0}, e_A, e_B, e_{AB}, Y_{MAT} \mid L_A, L_B, \Gamma_{AB}).$$

We simulate random draws from $P(\theta, Y_{MAT} \mid Y_{OBS})$ following the Markov chain Monte Carlo technique of data augmentation (Schafer, 1997, pg. 72, repeated here with notational changes):

> Given a current guess $\theta(t)$ of the parameter, first [the I-step] draw a value of the missing data from the conditional predictive distribution of Y_{MAT}, $Y_{MAT}(t+1) \sim P(Y_{MAT} \mid Y_{OBS}, \theta(t))$. Then [the P-step], conditional on $Y_{MAT}(t+1)$, draw a new value of θ from its complete-data posterior, $\theta(t+1) \sim P(\theta \mid Y_{OBS}, Y_{MAT}(t+1))$. Repeating this sampling from a starting value $\theta(0)$ yields a stochastic sequence $\{\theta(t), Y_{MAT}(t) : t = 1, 2, \ldots\}$ whose stationary distribution is $P(\theta, Y_{MAT} \mid Y_{OBS})$, and the subsequences $\{\theta(t) : t = 1, 2, \ldots\}$ and $\{Y_{MAT}(t) : t = 1, 2, \ldots\}$ have $P(\theta \mid Y_{OBS})$ and $P(Y_{MAT} \mid Y_{OBS})$ as their respective stationary distributions.

We use parallel chains from the same starting value $\theta(0)$ to generate multiple independent linkage imputations. $\theta(0)$ includes MLE values for p_A, p_B, e_{A0}, and e_{B0}.

4.2 Data Augmentation I-step

For Bayesian imputation of the true classification of pair (a,b), M or U, we again apply Bayes' rule for odds as shown in Section 3:

$$\mathrm{Odds}(Y_{MAT}(a,b) = 1 \mid \gamma(a,b), \theta(t)) = \prod_k (m_k(a,b) / u_k(a,b))\, P(M) / P(U),$$

where the product is over all comparison fields $k = 1, \ldots, K$ and the conditional probabilities $m_k(a,b)$ and $u_k(a,b)$ are calculated using the rules in Section 2. Note that the posterior odds and the likelihood ratio both depend on (a,b) but we assume that the prior odds do not.

As in the limited model, we choose an informative prior distribution for the odds based on substantive data. The limited model assumes a point estimate for the prior odds but for the full model we assume a lognormal distribution centered at the point estimate:

$$\log P(M) / P(U) \sim N(\log(N_{AB} / (N_A N_B - N_{AB})), \sigma^2).$$

We draw from the distribution for the prior odds once at the beginning of each I-step because the prior odds do not depend on (a,b).

Given $\mathrm{Odds}(Y_{MAT}(a,b) = 1 \mid \gamma(a,b), \theta(t))$, the posterior probability is

$$P(Y_{MAT}(a,b) = 1 \mid \gamma(a,b), \theta(t)) = \mathrm{Odds}(Y_{MAT}(a,b) = 1 \mid \gamma(a,b), \theta(t)) / 1 + \mathrm{Odds}(Y_{MAT}(a,b) = 1 \mid \gamma(a,b), \theta(t)).$$

We draw from $P(Y_{MAT} \mid Y_{OBS}, \theta(t)) = P(Y_{MAT} \mid \Gamma_{AB}, \theta(t))$ by drawing from a uniform random deviate X in (0,1) for each (a,b) and setting $Y_{MAT}(a,b) = 1$ if

$$P(Y_{MAT}(a,b) = 1 \mid \gamma(a,b), \theta(t)) > X.$$

4.3 Data Augmentation P-step

Fellegi and Sunter model data drawn from populations A, B, and A∩B as 3×K independent multinomial distributions with known parameters $p_A(k)$, $p_B(k)$, and $p_{AB}(k)$, $k = 1, \ldots, K$. The vector values parameter for each multinomial gives probabilities of observing each possible value of a comparison field in a sample from a population. For each field, nonresponse (missing values) and misreporting (incorrect values) are assumed to occur independently in the data capture process, completely at random. Probabilities of nonresponse ($e_{A0}(k)$ and $e_{B0}(k)$) and of misreporting ($e_A(k)$, $e_B(k)$, and $e_{AB}(k)$) are assumed to be known for all comparison fields $k = 1, \ldots, K$. These probabilities are assumed to be independent of field values.

In practice, $\theta = (p_A, p_B, p_{AB}, e_{A0}, e_{B0}, e_A, e_B, e_{AB})$ may not be known *a priori*—we only have independent samples from populations A and B, and, through the MCMC data augmentation procedure, from population A∩B produced in each I-step. All of the samples may include nonresponse and misreporting. Consequently, there is uncertainty about the true value of θ caused by sampling, nonresponse, and misreporting that should be modeled when drawing $\theta(t+1)$. In the Bayesian methodology described here, posterior distributions for all components of θ given Y_{OBS} and Y_{MAT} are independent because of the use of the latent class model and the assumption of prior independence of the components. All posterior distributions can be estimated using established techniques. Successive draws from the independent posterior distributions for the components of θ produce a draw from the full posterior distribution of θ.

4.3.1 Bayesian Models for p_A, p_B, e_{A0}, e_{B0}

We apply the same Bayesian analysis independently for each comparison field $k = 1, \ldots, K$ in sample L_A from population A and for each comparison field $k = 1, \ldots, K$ in sample L_B from population B. The approach here closely follows examples presented in Little and Rubin (2002, pp. 98–99, 114–115, and 120–121). The analysis is shown only for one field k in sample L_A but the analysis for other fields and samples is similar.

Denote field k in sample L_A as $L_A(k)$. Suppose $L_A(k) = (y_1, \ldots, y_{NA})^T$ where y_i is categorical and takes one of C possible values $c = 1, \ldots, C$. Let n_c be the number of observations for which $y_i = c$, with $\sum_c n_c = N_A$. Conditional on N_A, the counts $(n_1, \ldots, n_C)$ have a multinomial distribution with index N_A and probabilities $p_A(k) = (\pi_1, \ldots, \pi_C)$, $\pi_c > 0$, $\sum_c \pi_c = 1$. The likelihood is proportional to the distribution of $L_A(k)$ given $p_A(k)$

$$f(L_A(k) \mid p_A(k)) = (n! / \prod_c n_c!) \prod_c \pi_c^{nc}$$

For Bayesian inference, assume a Dirichlet prior distribution with vector parameter $\{\alpha_c\}$ for the parameters of the multinomial model:

$$P(\pi_1, \ldots, \pi_C) \propto \prod_c \pi_c^{\alpha c - 1}.$$

If a prior sample for field k from population A is available set α_c equal to the number of prior observations for which $y_i = c$ for all c. Otherwise, assume a proper non-informative prior distribution with $\alpha_c = 1$ for all c. The Dirichlet is a conjugate prior distribution for parameters of the multinomial model. Combining this prior distribution with the likelihood yields the posterior distribution as Dirichlet with vector parameter $\{n_c + \alpha_c\}$:

$$P(\pi_1, \ldots, \pi_C \mid L_A(k)) \propto \prod_c \pi_c^{nc + \alpha c - 1}.$$

Suppose the sample $L_A(k)$ is incomplete, observed for $O_A(k)$ units and missing for $M_A(k) = N_A - O_A(k)$ units. Denote the observed units in $L_A(k)$ as $L_{OBS}(k)$ and the missing units as $L_{MIS}(k)$ so that $L_A(k) = \{L_{OBS}(k), L_{MIS}(k)\}$. Let n_c be the number of observations for which $y_i = c$, with $\sum_c n_c = O_A(k)$. Conditional on $O_A(k)$, the counts $(n_1, \ldots, n_C)$ have a multinomial distribution with index $O_A(k)$ and probabilities $p_A(k) = (\pi_1, \ldots, \pi_C)$. The likelihood ignoring the missing-data mechanism is proportional to the distribution of $L_{OBS}(k)$ given $p_A(k)$, which is

$$f(L_{OBS}(k) \mid p_A(k)) = (O_A(k)! \,/\, \textstyle\prod_c n_c!) \prod_c \pi_c^{nc}.$$

Now consider the missing-data mechanism. Let $R(k) = (R_1, \ldots, R_{NA})^T$ measure nonresponse in sample $L_A(k)$, where $R_i = 0$ for observed units and $R_i = 1$ for missing units. We assume each unit in sample $L_A(k)$ is missing with probability $e_{A0}(k)$ independent of $L_A(k)$. Then the distribution of $R(k)$ given $L_A(k)$ and $e_{A0}(k)$ is

$$f(R(k) \mid L_A(k), e_{A0}(k)) = (N_A! \,/\, O_A(k)!\, M_A(k)!)\, (1 - e_{A0}(k))^{OA(k)}\, e_{A0}(k)^{MA(k)}.$$

The likelihood not ignoring the missing-data mechanism is proportional to the joint distribution for $L_{OBS}(k)$ and $R(k)$ given $p_A(k)$ and $e_{A0}(k)$

$$f(L_{OBS}(k), R(k) \mid p_A(k), e_{A0}(k)) = (N_A! \,/\, O_A(k)!\, M_A(k)!)\, (1 - e_{A0}(k))^{OA(k)}\, e_{A0}(k)^{MA(k)} \times (O_A(k)! \,/\, \textstyle\prod_c n_c!) \prod_c \pi_c^{nc}.$$

Missing data are Missing at Random (MAR) because we assume they are Missing Completely at Random (MCAR). Consequently, as shown by Little and Rubin (2002, pp. 120–121), if $p_A(k)$ and $e_{A0}(k)$ have independent priors $P(p_A(k))$ and $Q(e_{A0}(k))$ then Bayesian inferences about $p_A(k)$ can be based on $P(p_A(k))$ and the ignorable likelihood proportional to $f(L_{OBS} \mid p_A(k))$. The only effect of missing data is to decrease the effective sample size from N_A to $O_A(k)$. Bayesian inferences about $e_{A0}(k)$ given N_A and $O_A(k)$ can be based on $Q(e_{A0}(k))$ and the likelihood proportional to $f(R(k) \mid e_{A0}(k))$. With a binomial model for nonresponse and an independent Beta prior distribution for $e_{A0}(k)$ we obtain a Beta posterior distribution for $e_{A0}(k)$.

The sample $L_A(k)$ may contain misreported data. In general, any value can be misreported as any other value and a full treatment of misreported data is beyond the scope of this paper. We assume incorrect values in field k are indistinguishable from correct values and are drawn independently from the same distribution as correct values (same $p_A(k)$). In this case, inferences about $p_A(k)$ can be based on all cases in L_{OBS}, ignoring the fact that some values are incorrect. Note that if some incorrect values were distinguishable from correct values then we could remove those incorrect values for inferences about $p_A(k)$ with decreased effective sample size.

We simulate draws from Dirichlet or Beta posterior distributions using a related standard gamma distribution (Schafer, 1997, pg. 249), designated $G(a)$ with parameter $a > 0$. For each level c of a comparison field, simulate drawing v_c from $G(n_c + \alpha_c)$ where n_c is an observed count and α_c is a prior count (we take $\alpha_c = 1$ if prior counts are not available). Then $(v_1 / \sum v_c, v_2 / \sum v_c, \ldots, v_C / \sum v_c)$ is a simulated draw from the Dirichlet posterior with vector parameter $\{n_c + \alpha_c\}$.

4.3.2 Bayesian Models for p_{AB}, e_A, e_B, e_{AB}

We estimate the posterior distribution for $p_{AB}(k)$ given Y_{OBS} and $Y_{MAT}(t+1)$ by applying procedure in Section 4.3.1 independently for each comparison field $k = 1, \ldots, K$ in a sample L_{AB} from population $A \cap B$. The sample for each comparison field consists of those record pairs in $L_A \times L_B$ classified as matched in $Y_{MAT}(t+1)$ with agreements on the values of field k in Γ_{AB}.

We cannot estimate posterior distributions for $e_A(k)$, $e_B(k)$, and $e_{AB}(k)$ given Y_{OBS} and $Y_{MAT}(t+1)$ by directly analyzing those pairs $L_A \times L_B$ classified as matched in $Y_{MAT}(t+1)$ with disagreements on the values of field k in Γ_{AB}. When a comparison field k disagrees for some true matched pair one cannot determine by inspection whether the disagreement should be attributed to incorrect reporting in $L_A(k)$ with probability $e_A(k)$, incorrect reporting in $L_B(k)$ with probability $e_B(k)$, or correct but different reporting in $L_A(k)$ and $L_B(k)$ with probability $e_{AB}(k)$. For convenience, we let $e_T(k)$ be the consolidated probability of misreporting defined by

$$1 - e_T(k) = (1 - e_A(k))\,(1 - e_B(k))\,(1 - e_{AB}(k))$$

and estimate only the posterior distribution for $e_T(k)$ given Y_{OBS} and $Y_{MAT}(t+1)$. With our other assumptions, this has no important effect on calculated m and u probabilities because $e_A(k)$, $e_B(k)$, and $e_{AB}(k)$ occur only in the product $(1 - e_A(k))\,(1 - e_B(k))\,(1 - e_{AB}(k))$. We assume a binomial model for misreporting with an independent Beta prior distribution for $e_T(k)$ and obtain a Beta posterior distribution for $e_T(k)$.

4.3.3 Convergence of Posterior Distributions

Convergence to stationarity of posterior distributions of interest is not guaranteed when applying the MCMC data augmentation procedure. Recommended techniques for diagnosing convergence suggest comparing results from parallel chains with dispersed starting values (Gelman et al., 2000, Little and Rubin, 2004; Schafer, 1997). These techniques have not yet been implemented for this model. Instead, we monitor important summaries of the distributions by inspection. Given our assumptions, model parameters $p_A(k)$, $p_B(k)$, $e_{A0}(k)$, and $e_{B0}(k)$ describing comparison field k in populations A and B will be drawn from their respective stationary distributions after the first iteration. Of course, this may not be the case for model parameters $p_{AB}(k)$ and $e_T(k)$ describing comparison field k in population A∩B, or for Y_{MAT}, the true matched status of each record pair. We choose to monitor the combined error rate eT(k) for each comparison field k as well as N_M, the number of record pairs classified as true matches in Y_{MAT}. Preliminary test results using data from one CODES state suggest that convergence may occur quickly as shown in Table 2.

Table 2
Monitored Statistics for Assessing Convergence of Posterior Distributions

Monitored Statistic	***MCMC Iteration for Imputation 1*** $t=0$	$t=2$	$t=4$	$t=20$
eT(1)	0.19	0.12	0.12	0.12
eT(2)	0.19	0.20	0.21	0.22
eT(3)	0.19	0.05	0.04	0.04
eT(4)	0.19	0.09	0.06	0.06
eT(5)	0.19	0.10	0.07	0.07
eT(6)	0.19	0.03	0.02	0.02
eT(7)	0.19	0.17	0.16	0.16
N_M	8,000	9,676	9,184	9,140

4.3.3 Selecting Comparison Pairs

Only those record pairs in Γ_{AB} with agreement on at least one important comparison field are examined as comparison pairs because of practical limitations on computing time. Two or more independent match passes are performed, each joining files L_A and L_B on different fields to produce potentially different sets of comparison pairs. The union of unique comparison pairs from all passes is used to draw samples from A∩B. Only comparison pairs with posterior probabilities greater than 0.001 or some other low value established by worst-case analysis of the record generation process are included in the union. Pairs with lower posterior probabilities are assumed to be unmatched.

5. AREAS FOR FUTURE WORK

Linking Simulated Data. Different linkage modeling approaches will be compared by linking simulated data as in Fortini et al., (2001, 2002) and Larsen (1999, 2003).

Measuring Goodness of Fit. There is often a choice of alternate comparison variables for linkage models including variables with coarsened data. Area under ROC curves has been used as an overall measure of goodness of fit for logistic regression models and might be suitable for linkage models.

Modeling Misreporting Mechanisms. More realistic misreporting mechanisms will be modeled. Imputed samples of linked pairs and three-file links will be analyzed to estimate model parameters.

Varying Prior Odds for a Match. Newcombe (1995) suggests that prior odds for a true match might vary depending on personal characteristics. For example, an elderly driver might be more likely to be treated in a hospital than a young driver.

Expanding the Number of Linked Files. CODES analysts often link more than two files to build a full medical history for crash victims. The formulas in Section 2 will be expanded to cover links between three or more files with similar comparison outcomes. Candidate multiples will be found by conducting traditional pair-wise links.

Comparing Dependent Fields. Independence of comparison fields is measured by calculating uncertainty coefficients, information entropy based measures of association. When dependent comparison fields are used then their combined match weights for agreements will be reduced by the amount of common information.

ACKNOWLEDGEMENTS

I am grateful to Sandra Johnson, independent consultant to NHTSA, Jack Leiss and Ming Yin at Constella Health Sciences, and especially Michael Larsen at Iowa State University for suggestions which substantially improved this paper. The research was funded in part by the Governors' Highway Safety Association.

REFERENCES

BELIN, T.R. and RUBIN, D.B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, **90**, 694–707.

FELLEGI, I.P. and SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, **64**, 1183–1210.

FORTINI, M., LISEO, B., NUCCITELLI, A., SCANU, M. (2001). On Bayesian record linage. *Research in Official Statistics*, **4**, 184–198.

FORTINI, M., NUCCITELLI, A., SCANU, M., LISEO, B. (2002). Modelling issues in record linkage: A Bayesian perspective. *Proceedings of the American Statistical Association Meeting, August 2002.*

GELMAN, A., CARLIN, J.B., STERN, H.S., and RUBIN, D.B. (1995). *Bayesian Data Analysis*. Chapman & Hall/CRC.

GREENBERG, L. (1996). *Police Accident Report (PAR) Quality Assessment Project*. Technical Report DOT HS 808 487, National Highway Traffic Safety Administration.

GUSTAFSON, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Chapman & Hall/CRC.

JARO, M.A. (1992). *AUTOMATCH Generalized Record Linkage System User's Manual*. MatchWare Technologies, Inc.

JARO, M.A. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine*, **14**, 491–498.

LARSEN, M.D. (1999). Multiple imputation analysis of records linked using mixture models. *Proceedings of the Survey Methods Section, Statistics Society of Canada Annual Meeting, June 1999*, 65–71.

LARSEN, M.D. (2003). Hierarchical Bayesian record linkage and regression in linked files. *Joint Summer Research Conference on Machine Learning, Statistics, and Discovery; AMS, IMS, SIAM, June 2003*.

LARSEN, M.D. and RUBIN, D.B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, **96**, 32–41.

LITTLE, R.J.A. and RUBIN, D.B (2002). *Statistical Analysis with Missing Data* (2nd edition). Wiley.

MCGLINCY, M.H. (2003). *Touring CODES2000 and LinkSolv Applications*. Strategic Matching, Inc. Technical Report.

MCGLINCY, M.H., GUARDINO, J., and Associates (1994). *New York State Crash Outcome Data Evaluation System (CODES) Project—1992 Final Report*. Albany: New York State Department of Health.

NEWCOMBE, H.B. (1995). Age-related bias in probabilistic death searches due to neglect of the "prior likelihoods." *Computers and Biomedical Research*, **28**, 87–99.

RUBIN, D.B., SCHAFER, J.L., and SUBRAMANIAN, R. (1998). *Multiple Imputation of Missing Blood Alcohol Concentration (BAC) Values in FARS*. Technical Report DOT HS 808 816, National Highway Traffic Safety Administration.

RUNGE, J.W. (2000). Linking data for injury control research. *Annals of Emergency Medicine*, **35**, 613–615.

SCHAFER, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Boca Raton: Chapman & Hall/CRC.

SCHEUREN, F. and WINKLER, W. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, **19**, 39–58.

SCHEUREN, F. and WINKLER, W. (1997). Regression analysis of data files that are computer matched – Part II. *Survey Methodology*, **23**, 157–165.

VERNON, D.D., COOK, L.J., PETERSON, K.J. and DEAN, J.M. (2004). Effect of repeal of the national maximum speed limit law on occurrence of crashes, injury crashes, and fatal crashes on Utah highways. *Accident Analysis & Prevention*, ***36***, 223–229.

WINKLER, W. E. (1988). Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 667– 671.

WINKLER, W. E. (1989). Frequency-based matching in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 778–783.

WINKLER, W. E. (1993). Improved decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 274–279.

WINKLER, W. E. (1994). Advanced methods for record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 467–472.